STAT 300 Chapter 5

Zepu Zhang
March 22, 2012

There are three main themes in this chapter:
(1) Various aspects of multivariate distributions;
(2) Interactions between two variables in a joint distribution;
(3) Distribution of functions (or combinations) of random variables.

1 Joint distributions

Univariate versus multivariate.

Multivariate random variable a.k.a. "random vector".

1.1 Discrete: pmf

The distribution of a univariate discrete RV X is specified by the probability of every value of the variable:

$$p(x) = P(X = x)$$

where $p(x) \geq 0$ and $\sum_x p(x) = 1$.

The distribution of a bivariate discrete RV (X, Y) is specified by the probability of every pair of values of X and Y, or every value of the vector (X, Y):

$$p(x, y) = P(X = x \text{ and } Y = y)$$

where $p(x, y) \geq 0$ and $\sum_x \sum_y p(x, y) = 1$.

We call this either the "distribution of a bivariate random variable" or the "joint distribution of two (univariate) random variables".

Example Ex. 5.1 in Chap. 5.1. X: car insurance deductible; Y: home insurance deductible.

		y		
	$p(x,y)$	0	100	200
x	100	.20	.10	.20
	250	.05	.15	.30

Exercise Verify this specifies a valid joint distribution. Hint: check two things: (1) all p's are ≥ 0; (2) all p's sum to 1.

Generalization to more than two variables: straightforward.

1.2 Continuous: pdf

The distribution of a univariate continuous RV X is specified by a density function $f(x)$ such that

$$P(a < X < b) = \int_a^b f(x)\,\mathrm{d}x$$

where $f(x) \geq 0$ and $\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1$ (the latter is implied by the relation above).

Visualization: a curve; area under it.

> Remember it's not useful to talk about the probability that X takes a particular value, because $P(X = x) = 0$ for any particular value x. Hence we turn to an integral on an interval.

The distribution of a bivariate continuous RV (X, Y) is specified by a density function $f(x, y)$ such that

$$P\big[(X, Y) \in A\big] = \iint\limits_A f(x, y)\,\mathrm{d}x\,\mathrm{d}y$$

where $f(x, y) \geq 0$ and $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)\,\mathrm{d}x\,\mathrm{d}y = 1$; A is a 2-D set. For example, $A = \{(x, y) : a \leq x \leq b, c \leq y \leq d\}$.

> A is a contiguous region, or union of such regions. It's not a set of discrete points. Again, $P\big((X, Y) = (x, y)\big) = 0$ for any specific value (x, y).

Visualization: a surface; volume under it.

Note Following the concept above, calculating probabilities involves integration, e.g.
$P(a < X < b, c < Y < d) = \int_c^d \int_a^b f(x, y)\,\mathrm{d}x\,\mathrm{d}y$,
$P(a < X < b) = \int_{-\infty}^{\infty} \int_a^b f(x, y)\,\mathrm{d}x\,\mathrm{d}y$,
$P(c < Y < d) = \int_c^d \int_{-\infty}^{\infty} f(x, y)\,\mathrm{d}x\,\mathrm{d}y$,
$P(X > 3) = \int_{-\infty}^{\infty} \int_3^{\infty} f(x, y)\,\mathrm{d}x\,\mathrm{d}y$.

Example Ex. 5.3 in Chap. 5.1. X: proportion of time the drive-up is busy; Y: proportion of time the walk-up is busy.

$$f(x, y) = \begin{cases} \frac{6}{5}(x + y^2), & 0 \le x \le 1,\ 0 \le y \le 1 \\ 0, & \text{otherwise} \end{cases}$$

Exercise Verify this specifies a valid distribution. Hint: check two things: (1) $f(x, y) \ge 0$; (2) $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)\,\mathrm{d}x\,\mathrm{d}y = 1$.

Generalization to more than two variables: natural, but hard to visualize.

2 Expected values

Discrete:

$$E[h(X)] = \sum_x h(x)\, p(x)$$
$$E[h(X, Y)] = \sum_x \sum_y h(x, y)\, p(x, y)$$

Continuous:

$$E[h(X)] = \int_{-\infty}^{\infty} h(x) f(x)\,\mathrm{d}x$$

$$E[h(X, Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x, y) f(x, y)\,\mathrm{d}x\,\mathrm{d}y$$

Note, $h(X, Y)$ is a <u>function</u> that evaluates to <u>one value</u>. While for univariate we can define $h(X) = X$ and get the expected value of X, there is no way to define $h(X, Y) = (X, Y)$ and get the expected value of the variable <u>pair</u> at once.

However, it's no problem to define $h(X, Y) = X$. Then

$$E[X] = \sum_x \sum_y x \cdot p(x, y)$$

$$E[X] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x \cdot f(x, y)\,\mathrm{d}x\,\mathrm{d}y$$

Similarly,

$$E[Y] = \sum_x \sum_y y \cdot p(x, y)$$

$$E[Y] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y \cdot f(x,y)\,\mathrm{d}x\,\mathrm{d}y$$

Then we may say

$$E\big[(X,Y)\big] = \big(E(X), E(Y)\big)$$

Example Ex. 5.14 in Chap. 5.2. (Optional)

3 Covariance and correlation

If we consider X, ignoring Y, the variance of X is defined as before, that is, take $h(X,Y) = \big(X - E(X)\big)^2$ and define

$$\operatorname{var}(X) = E\big(h(X,Y)\big) = \sum_x \sum_y \big(x - E[X]\big)^2 \cdot p(x,y)$$

in the discrete case and

$$\operatorname{var}(X) = E\big(h(X,Y)\big) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \big(x - E[X]\big)^2 \cdot f(x,y)\,\mathrm{d}x\,\mathrm{d}y$$

in the continuous case.

However, our interest now is more in the "joint" behavior of X and Y. The most important summary measure of the "association" between X and Y (that is, relation like "Y tends to be big when X is big" or the other way around) is the covariance or correlation.

3.1 Covariance

Covariance describes the "association" between X and Y (in a statistical sense, as both vary randomly).

Definition Covariance:

$$\operatorname{cov}(X,Y) = E\big[(X - \mu_X)(Y - \mu_Y)\big]$$

Rational: p. 198 (7th ed) or p. 208 (8th ed).

A useful formula

$$\operatorname{cov}(X,Y) = E(XY) - E(X)\,E(Y)$$

Exercise Prove the equality above. Hint:

$$\operatorname{cov}(X,Y) = E\big[XY - X\mu_Y - \mu_X Y + \mu_X\mu_Y\big] = E[XY] - E[X\mu_Y] - E[\mu_X Y] + E[\mu_X\mu_Y] = \cdots$$

Remark In the univariate case we learned the definition $\operatorname{var}(X) = E\big[(X-\mu_X)^2\big]$ and the relation $\operatorname{var}(X) = E(X^2) - \mu_X^2$. Compare them with the definition of and relation about covariance above. In particular, if you take $\operatorname{var}(X)$ to be $\operatorname{cov}(X,X)$ and use the definition of and relation about covariance above, what do you get?

Example Ex. 5.16 in Chap. 5.2. (Optional)

$$f(x,y) = \begin{cases} 24xy, & 0 \le x \le 1,\ 0 \le y \le 1,\ x+y \le 1 \\ 0, & \text{otherwise} \end{cases}$$

3.2 Correlation coefficient

One can tell the "direction" (positive or negative) of the association based on $\operatorname{cov}(X,Y)$. But it is hard to tell the <u>strength</u> of the association from the magnitude of $\operatorname{cov}(X,Y)$.

To eliminate this difficulty, we strive to "standardize" $\operatorname{cov}(X,Y)$ so that its value is restricted to a certain range, and then we can tell the strength of the association by observing whether the value is near the limits. This "standardized" version of covariance is called the "correlation coefficient".

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \cdot \sigma_Y}$$

Remarks

1. Sometimes written as $\operatorname{cor}(X,Y)$ or $\operatorname{corr}(X,Y)$.

2. $-1 \le \rho \le 1$.

3. ρ does not change with linear transform, that is, $\operatorname{cor}(aX+b, cY+d) = \operatorname{cor}(X,Y)$.

   To prove this property, notice $\operatorname{cov}(aX+b, cY+d) = \operatorname{cov}(aX, cY) = ac\operatorname{cov}(X,Y)$ and $\operatorname{var}(aX+b) = a^2\operatorname{var}(X)$, $\operatorname{var}(cY+d) = c^2\operatorname{var}(Y)$.

4. Empirical interpretation: $|\rho| > .8$: strong correlation; $.5 < |\rho| < .8$: moderate correlation; $|\rho| < .5$: weak correlation.

5. Correlation coefficient describes linear association only. "No linear association" can be totally different from "no association".

6. Association does not establish causality.

7. When $\rho = 0$ we say X and Y are uncorrelated. Hence being uncorrelated means no linear relation.

8. "Uncorrelated" is different from "independent". Being "uncorrelated" does not mean "un-related"!

3.3 Independence

Without introducing the definition of independence between two random variables, we make a few points. (We learned the definition of independence between two events.)

If X and Y are independent, then

1. $$P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$$

   In words, events about A and events about B are independent.

2. $$E(XY) = E(X) \cdot E(Y)$$

3. and, consequently,

   $$\mathrm{cov}(X, Y) = 0$$

   and

   $$\rho(X, Y) = 0$$

4. But, the converse is not true. $\mathrm{cov}(X, Y) = 0$ does not prove independence between X and Y.

4 Distribution of a linear combination

4.1 Linear combination of arbitrary RV

Let $X_1, \ldots, X_n$ be different RV (meaning "random variables") with means $\mu_1, \ldots, \mu_n$ and variances $\sigma_1^2, \ldots, \sigma_n^2$.

Note on notation: $X_1, \ldots, X_n$ are random variables. They do not need to have any specific relation. Don't be confused by the common X in the symbols.

Let $Y = \sum_{i=1}^{n} a_i X_i$, where a_i are constants. Note that Y is a univariate RV.

Proposition

$$E(Y) = \sum_{i=1}^{n} E(a_i X_i) = \sum_{i=1}^{n} a_i E(X_i)$$

$$\operatorname{var}(Y) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \operatorname{cov}(X_i, X_j)$$

Note The double summation above contains terms where $i = j$; in those terms $a_i a_j \operatorname{cov}(X_i, X_j)$ is just $a_i^2 \operatorname{var}(X_i)$.

Special case: if X_1,..., X_n are independent, then $\operatorname{cov}(X_i, X_j) = 0$ for $i \neq j$, hence

$$\operatorname{var}(Y) = \sum_{i=1}^{n} a_i^2 \operatorname{var}(X_i)$$

Since the constants a_i can be negative, the statements above include situations of "subtractions".

Proof For the case $n = 2$: end of Chap. 5.5.

4.2 Linear combination of normal RV

The preceding statements are still correct, of course.

In fact we know something much stronger: Y is normal. In words, a linear combination of normal random variables is normal (whether the variables are independent or not).

As a result $E(Y)$ and $\operatorname{var}(Y)$ completely describe the distribution of Y.

5 Statistics and sampling distributions

Definition Statistic: a function of sample data.

The sample data are uncertain prior to the actual sampling, and their values vary as the same sampling procedure is repeated. Consequently, the value of a statistic is also random. Therefore, a statistic is a RV.

Example Sample mean, $\overline{X} = \frac{1}{n}(X_1 + \dots X_n)$, is a statistic. Hence it's a RV and has a distribution, known as the sampling distribution (which describes the variation of the statistic during sampling).

Definition The RV's $X_1, \dots, X_n$ form a simple random sample if
1. The X_i's are independent;
2. Every X_i has the same probability distribution.

This is called an "iid" (independent, identically distributed) sample.

Read: paragraph after definition box in section "Random Samples", Chap. 5.3.

Nice explanations on notation: first paragraph of Chap. 5.3; paragraph below definition box for "statistic" in first section of Chap. 5.3.

Example Ex. 5.22 and 5.23 in Chap. 5.3. (Self reading.)

6 Distribution of sample mean

$$\overline{X} = \frac{X_1 + \dots X_n}{n}$$

where X_i's are iid with mean μ and variance σ^2.

Since this is a linear combination of RV's, we've learned

$$\begin{aligned} E\left(\overline{X}\right) &= \frac{E(X_1) + \dots + E(X_n)}{n} = \mu \\ \operatorname{var}\left(\overline{X}\right) &= \frac{\operatorname{var}(X_1) + \dots + \operatorname{var}(X_n)}{n^2} = \sigma^2/n \quad \text{(using "independence")} \end{aligned}$$

In addition, if $X_i \sim N(\mu, \sigma^2)$, then

$$\overline{X} \sim N(\mu, \sigma^2/n)$$

The sample mean is an important statistic. It is used for estimating the population mean, μ.

Example Ex. 5.25 in Chap. 5.4.

7 Central limit theorem (CLT)

Theorem Let $X_1, \ldots, X_n$ be an iid sample from a distribution with mean μ and variance σ^2. Then we already know $E(\overline{X}) = \mu$ and $\text{var}(\overline{X}) = \sigma^2/n$. Moreover, we know that $\overline{X}$ has the normal distribution, $N(\mu, \sigma^2/n)$, if the X_i's are normal. Now if the X_i's are not normal, CLT states that $\overline{X}$ is approximately normal when n is "sufficiently large", regardless of the original distribution of the X_i's (it can be continuous or discrete, symmetric or skewed, close to normal or rather different). In symbols,

$$\overline{X} \longrightarrow N(\mu, \sigma^2/n) \quad \text{and} \quad \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \longrightarrow N(0, 1), \qquad \text{as } n \to \infty$$

Note 1. n need not be huge. Rule of thumb: $n > 30$ is usually large enough.

2. The quality of the normal approximation depends on how severely the original distribution differs from normal, and how large n is.

3. The theorem requires X_i's to be independent and identically distributed.

4. This theorem is the basis of why the normal distribution is so commonly applicable, and so fundamental.

5. Note that if the X_i's are normal, we don't need to invoke CLT. In that case we know $\overline{X}$ is (exactly, not approximately) normal however large or small n is; moreover, the X_i's do not need to be independent, and their normal distributions do not need to be identical.

Example Ex. 5.26 in Chap. 5.4.

Exercise Compare Ex. 5.26 and Ex. 5.25. What are the differences in the given conditions?